O EFEITO BACKFIRE GERADO PELA INTELIGÊNCIA ARTIFICIAL EM SUA EPISTEMOLOGIA, COGNIÇÃO E OS PARADOXOS DA CORREÇÃO INFORMACIONAL

THE BACKFIRE EFFECT GENERATED BY ARTIFICIAL INTELLIGENCE IN ITS EPISTEMOLOGY, COGNITION, AND THE PARADOXES OF INFORMATIONAL CORRECTION

Atila Barros da Silva[1]
DOI: 10.66029/UJPP4043

RESUMO: O presente artigo empreende uma análise aprofundada do efeito backfire, fenômeno pelo qual tentativas de corrigir informações falsas ou crenças errôneas resultam, paradoxalmente, no fortalecimento delas, à luz das transformações epistemológicas e cognitivas introduzidas pelas novas tecnologias de inteligência artificial. Partindo da formulação seminal de Nyhan e Reifler (2010), o texto articula o fenômeno com três eixos analíticos: a arqueologia do saber proposta por Foucault (1999, 2008), que desvela as condições históricas de possibilidade da verdade; a teoria da autodeterminação de Deci e Ryan (1985, 2000, 2017), que acende as motivações psicológicas subjacentes à resistência à correção; e as reflexões de Rovelli (2017, 2018) sobre a natureza temporal e perspectiva da realidade, que questionam a própria noção de "fato" objetivo. Argumenta-se que a IA, ao mesmo tempo que amplia exponencialmente a capacidade de geração e disseminação de desinformação, também introduz novos vetores para o efeito backfire, desde a percepção de viés algorítmico até a persuasão generativa enganosa, ao passo que oferece ferramentas inéditas para sua mitigação. A análise culmina em uma reflexão crítica sobre os limites éticos e epistemológicos das estratégias de correção baseadas em IA, à luz do problema da simulação proposto por Bostrom (2003) e das implicações existenciais da superinteligência (Bostrom, 2014).

Palavras-Chave: Efeito backfire. Inteligência Artificial. Desinformação. Teoria da Autodeterminação

ABSTRACT: This article undertakes an in-depth analysis of the backfire effect, a phenomenon whereby attempts to correct false information or erroneous beliefs paradoxically result in their strengthening, in light of the epistemological and cognitive transformations introduced by new artificial intelligence

[1] Professor e pesquisador de Tecnologias Digitais, Pedagogia, Sociologia e Antropologia, Universidade Estácio de Sá. Lattes: http://lattes.cnpq.br/4660184868094532. Orcid iD: https://orcid.org/0009-0006-3739-0681. E-mail: atilafmusp@gmail.com

REVISTA MOVER — Ciências e Humanidades | Revista Acadêmica de Acesso Aberto.
©2026 POR MOVER
REVISTAMOVER.COM.BR

technologies. Drawing on the seminal formulation of Nyhan and Reifler (2010), the text articulates the phenomenon through three analytical axes: the archaeology of knowledge proposed by Foucault (1999, 2008), which unveils the historical conditions of possibility for truth; the self-determination theory of Deci and Ryan (1985, 2000, 2017), which illuminates the psychological motivations underlying resistance to correction; and the reflections of Rovelli (2017, 2018) on the temporal and perspectival nature of reality, which call into question the very notion of an objective "fact." It is argued that AI, while exponentially amplifying the capacity for generating and disseminating misinformation, also introduces new vectors for the backfire effect, ranging from the perception of algorithmic bias to deceptive generative persuasion, while concurrently offering unprecedented tools for its mitigation. The analysis culminates in a critical reflection on the ethical and epistemological limits of AI-based correction strategies, in light of the simulation problem proposed by Bostrom (2003) and the existential implications of superintelligence (Bostrom, 2014).

Keywords: Backfire effect. Artificial Intelligence. Misinformation. Self-Determination Theory.

RESUMEN: El presente artículo emprende un análisis profundo del efecto backfire, fenómeno por el cual los intentos de corregir informaciones falsas o creencias erróneas resultan, paradójicamente, en su fortalecimiento, a la luz de las transformaciones epistemológicas y cognitivas introducidas por las nuevas tecnologías de inteligencia artificial. Partiendo de la formulación seminal de Nyhan y Reifler (2010), el texto articula el fenómeno con tres ejes analíticos: la arqueología del saber propuesta por Foucault (1999, 2008), que devela las condiciones históricas de posibilidad de la verdad; la teoría de la autodeterminación de Deci y Ryan (1985, 2000, 2017), que ilumina las motivaciones psicológicas subyacentes a la resistencia a la corrección; y las reflexiones de Rovelli (2017, 2018) sobre la naturaleza temporal y perspectiva de la realidad, que cuestionan la propia noción de "hecho" objetivo. Se argumenta que la IA, al mismo tiempo que amplía exponencialmente la capacidad de generación y diseminación de desinformación, también introduce nuevos vectores para el efecto backfire, desde la percepción de sesgo algorítmico hasta la persuasión generativa engañosa, al tiempo que ofrece herramientas inéditas para su mitigación. El análisis culmina en una reflexión crítica sobre los límites éticos y epistemológicos de las estrategias de corrección basadas en IA, a la luz del problema de la simulación propuesto por Bostrom (2003) y de las implicaciones existenciales de la superinteligencia (Bostrom, 2014).

©2026 POR MOVER

Palabras Clave: Efecto backfire. Inteligencia Artificial. Desinformación. Teoría de la Autodeterminación.

REVISTA MOVER — Ciências e Humanidades | Revista Acadêmica de Acesso Aberto.
©2026 POR MOVER
REVISTAMOVER.COM.BR

INTRODUÇÃO

O novo mundo digital, caracterizada pela ubiquidade das informações e pela velocidade vertiginosa de sua circulação, instituiu um paradoxo central para as democracias contemporâneas: nunca houve tanto acesso ao conhecimento, e nunca a crença no falso foi tão persistente e socialmente disruptiva. No coração desse paradoxo reside o efeito backfire, fenômeno documentado empiricamente por Brendan Nyhan e Jason Reifler (2010), segundo o qual a exposição a correções factuais pode, em determinados contextos, aumentar a adesão do indivíduo à crença incorreta que se pretende corrigir. Longe de ser uma anomalia cognitiva marginal, o efeito backfire revela dimensões profundas da psicologia humana, da dinâmica social e da própria constituição do sujeito epistêmico na modernidade tardia.

A emergência e a rápida disseminação das tecnologias de inteligência artificial (IA), especialmente dos grandes modelos de linguagem (*large language models*, LLMs) e dos sistemas generativos, impõem uma reavaliação urgente desse fenômeno. Se, por um lado, a IA promete revolucionar os mecanismos de verificação factual e de combate à desinformação, por outro, ela própria se constitui como uma fonte inédita de geração de conteúdo enganoso, capaz de produzir explicações logicamente articuladas e persuasivas que amplificam crenças falsas (Danry et al., 2024). Mais ainda, a mediação algorítmica da informação, com seus sistemas de recomendação personalizados e a consequente formação de bolhas informacionais e câmaras de eco (Pariser, 2011), cria um ambiente propício à cristalização de vieses cognitivos e à polarização social.

A abordagem aqui empreendida distancia-se deliberadamente das leituras reducionistas que circunscrevem o efeito backfire ao domínio da psicologia cognitiva ou do condicionamento comportamental, para inseri-lo em um quadro interpretativo de alcance epistemológico, motivacional e ontológico. Tal esforço exige a mobilização de três horizontes teóricos que, embora heterogêneos em suas filiações disciplinares, convergem para iluminar facetas complementares do fenômeno. O primeiro desses horizontes é fornecido pela arqueologia do saber de Michel Foucault, tal como desenvolvida em As palavras e as coisas (1999), e retomada em A coragem da verdade (2008). A obra foucaultiana permite compreender que a eficácia performática da correção factual não é um atributo intrínseco da proposição verdadeira, mas depende de um regime de verdade historicamente constituído, uma episteme, no qual a correspondência entre enunciado e realidade objetiva

REVISTA MOVER — Ciências e Humanidades | Revista Acadêmica de Acesso Aberto.
©2026 POR MOVER
REVISTAMOVER.COM.BR

funciona como critério dominante de validação. Quando esse regime se fragmenta, como ocorre em contextos de polarização discursiva e de multiplicação de autoridades epistêmicas concorrentes, a correção perde sua força vinculante e pode, paradoxalmente, reforçar a adesão ao enunciado falso, na medida em que este passa a operar como marcador de identidade e como resistência a uma pretensa ortodoxia imposta de fora. O segundo eixo teórico, de natureza psicológico-motivacional, é extraído da teoria da autodeterminação de Edward Deci e Richard Ryan (1985, 2000, 2017), cuja contribuição central reside na identificação de três necessidades psicológicas inatas, autonomia, competência e pertencimento, cuja satisfação é condição para o florescimento humano e para a motivação autônoma. Sob essa lente, a correção externa de uma crença arraigada não se reduz a um mero confronto lógico entre proposições; ela é vivenciada como uma ameaça existencial à integridade do self, na medida em que põe em questão a capacidade de julgar com independência (autonomia), a posse de conhecimentos corretos (competência) e a sintonia com o grupo de referência (pertencimento). O efeito backfire emerge, assim, como um mecanismo de defesa psicologicamente funcional, ainda que epistemicamente disfuncional, cuja força reside na restauração do equilíbrio homeostático dessas necessidades fundamentais. O terceiro aporte teórico, de caráter mais especulativo e ontológico, provém das reflexões de Carlo Rovelli (2017, 2018) sobre a natureza relacional e perspectiva da realidade física, bem como das hipóteses de Nick Bostrom (2003, 2014), acerca da simulação computacional e dos destinos da inteligência em escala cósmica. Rovelli, ao demonstrar que o tempo e a própria materialidade do mundo são propriedades emergentes de interações entre sistemas, e não dados absolutos, convida a uma desconfiança radical em relação a qualquer noção de "fato" como entidade atômica e independente do observador. Essa desconfiança ganha contornos práticos quando transposta para o ecossistema informacional contemporâneo, no qual a IA gera conteúdos que se adaptam dinamicamente ao perfil de cada usuário, turvando a distinção entre representação fiel e construção persuasiva da realidade. Bostrom, por sua vez, ao especular que nossa própria realidade pode ser uma simulação e ao alertar para os riscos de uma superinteligência capaz de manipular sistemas de crença com precisão inimaginável, eleva a questão a um patamar existencial: se o próprio substrato do real é passível de simulação, a correção factual perde seu fundamento metafísico e se torna uma disputa entre mundos possíveis, cada um com sua coerência interna e seus critérios de veracidade. Longe de constituírem perspectivas excludentes, esses três olhares, o arqueológico-discursivo, o motivacional-existencial e o ontológico-simulacional, entrelaçam-se para compor um modelo interpretativo do efeito backfire que recusa tanto o individualismo metodológico

REVISTA MOVER — Ciências e Humanidades | Revista Acadêmica de Acesso Aberto.
©2026 POR MOVER
REVISTAMOVER.COM.BR

quanto o determinismo tecnológico, e que insiste na natureza multifacetada e historicamente situada do fenômeno.

A tese central que emerge dessa arquitetura teórica é a de que o efeito backfire não pode ser adequadamente compreendido como um simples viés cognitivo ou uma anomalia da racionalidade instrumental, mas antes como um sintoma profundo, e talvez inevitável, das transformações epistemológicas que atravessam a cultura ocidental contemporânea. As tecnologias de inteligência artificial, longe de serem meros instrumentos neutros de processamento de informação, atuam como catalisadores e espelhos dessas transformações: elas refletem a fragmentação dos regimes de verdade ao mesmo tempo que a intensificam, ao gerarem explicações enganosas com plausibilidade lógica, ao personalizarem bolhas informacionais que blindam o usuário contra o contraditório e ao suscitarem desconfiança em relação à própria mediação algorítmica. Contudo, essa mesma potência disruptiva contém em si a possibilidade de uma reconfiguração responsável das práticas corretivas, desde que estas sejam concebidas não como imposições unilaterais de um fato supostamente autoevidente, mas como intervenções dialógicas que reconheçam a legitimidade das necessidades motivacionais do interlocutor e a historicidade de suas crenças. A compreensão do efeito backfire em sua complexidade integral, epistemológica, psicológica e ontológica, não é, portanto, um luxo acadêmico, mas uma condição prática e ética inescapável para o desenho de estratégias de combate à desinformação que sejam ao mesmo tempo eficazes, respeitosas da autonomia dos sujeitos e conscientes das limitações inerentes a qualquer pretensão de objetividade num mundo mediado por simulações algorítmicas.

O EFEITO BACKFIRE ENTRE A COGNIÇÃO E A EPISTEMOLOGIA

O efeito backfire, tal como formulado por Nyhan e Reifler (2010), designa a ocorrência, em contextos experimentalmente controlados, de um aumento das percepções errôneas entre indivíduos de um determinado grupo ideológico após a apresentação de uma correção factual. Em seus experimentos, os autores verificaram que correções de alegações políticas enganosas não apenas falhavam em reduzir as crenças incorretas entre os membros do grupo-alvo, mas, em alguns casos, produziam o efeito oposto, consolidando ainda mais a adesão à informação falsa. Esse achado, que desafiou o pressuposto iluminista de que a informação correta é suficiente para desfazer o erro, inaugurou um campo de investigação sobre os limites da racionalidade humana diante da dissonância cognitiva e da identidade grupal.

REVISTA MOVER — Ciências e Humanidades | Revista Acadêmica de Acesso Aberto.
©2026 POR MOVER
REVISTAMOVER.COM.BR

A literatura subsequente, no entanto, revelou que o efeito backfire não é um fenômeno universal ou inevitável. Estudos de replicação e meta-análises sugerem que sua ocorrência depende de uma constelação de fatores contextuais, incluindo o grau de polarização do tema, a força da identidade partidária do indivíduo, o tom e a fonte da correção, e a presença de contra-argumentos afetivamente carregados. Essa complexidade aponta para a necessidade de uma compreensão mais profunda dos mecanismos psicológicos e sociais subjacentes ao fenômeno.

Neste ponto, a teoria da autodeterminação (TAD) de Deci e Ryan oferece um arcabouço explicativo particularmente fecundo. A TAD postula a existência de três necessidades psicológicas inatas, autonomia, competência e pertencimento, cuja satisfação é essencial para o bem-estar e para a motivação intrínseca (Deci & Ryan, 1985, 2000; Ryan & Deci, 2017). A correção de uma crença central para a identidade do indivíduo pode ser percebida como uma ameaça à sua autonomia (a sensação de que está sendo manipulado ou coagido), à sua competência (a admissão de que estava errado) e ao seu pertencimento (o risco de isolamento do grupo de referência). O efeito backfire emergiria, então, como uma estratégia de defesa psicológica para restaurar o equilíbrio dessas necessidades ameaçadas: ao reforçar a crença original, o indivíduo reafirma sua autonomia, protege sua autoestima e mantém sua conexão com o grupo.

Essa leitura psicológica encontra eco e aprofundamento na obra de Michel Foucault, particularmente em As palavras e as coisas (1999) e em A coragem da verdade (2008). Foucault nos ensina que a verdade não é uma entidade transcendente, mas o produto de práticas discursivas e de regimes de poder-saber que variam historicamente. Em *As palavras e as coisas*, ele demonstra como as epistemes, as configurações históricas que definem as condições de possibilidade do conhecimento, se transformam ao longo do tempo, deslocando o que pode ser dito, pensado e reconhecido como verdadeiro (Foucault, 1999). Nessa perspectiva, a correção factual, tal como praticada pelo fact-checking tradicional, pressupõe uma episteme na qual a correspondência entre proposição e realidade objetiva é o critério último de verdade. No entanto, em contextos de polarização intensa e de fragmentação discursiva, essa episteme pode entrar em colapso, dando lugar a regimes de verdade concorrentes, nos quais a autoridade epistêmica não reside mais na "adequação à realidade", mas na fidelidade ao grupo, na coerência narrativa ou na performatividade emocional.

A coragem da verdade (Foucault, 2008), por sua vez, explora a noção de *parrhesia*, o discurso verdadeiro que envolve coragem e

REVISTA MOVER — Ciências e Humanidades | Revista Acadêmica de Acesso Aberto.
©2026 POR MOVER
REVISTAMOVER.COM.BR

risco por parte de quem o enuncia. Foucault mostra que o dizer-verdadeiro não é um ato neutro, mas uma prática ética e política que implica uma relação específica consigo mesmo e com os outros. O efeito backfire pode ser interpretado como uma recusa à *parrhesia* alheia: a correção factual é percebida não como um gesto de cuidado ou de busca compartilhada pela verdade, mas como uma imposição autoritária, um ato de violência simbólica que desrespeita a autonomia do sujeito e sua capacidade de autodeterminação. Nesse sentido, a resistência à correção não é mera irracionalidade, mas uma forma de luta pela própria verdade, uma verdade que é constitutiva da identidade e da dignidade do sujeito.

As reflexões de Carlo Rovelli sobre a natureza do tempo e da realidade acrescentam uma camada adicional de complexidade. Em A ordem do tempo (2018) e em A realidade não é o que parece (2017), Rovelli argumenta que a realidade física é fundamentalmente relacional e perspectiva, e que o tempo, longe de ser uma estrutura absoluta, emerge das interações entre sistemas. Transpondo essas ideias para o domínio da informação, podemos questionar se a noção de "fato" objetivo, que fundamenta a prática da correção, não é ela própria uma simplificação grosseira de uma realidade infinitamente mais complexa e perspectiva. Em um mundo onde a IA gera conteúdos que se adaptam e se personalizam para cada usuário, a própria ideia de uma "versão correta" dos fatos torna-se problemática. O efeito backfire, nessa leitura, seria uma manifestação da resistência a uma pretensão de objetividade que a própria física contemporânea nos ensina a desconfiar.

A IA COMO AMPLIFICADORA DO VIÉS ALGORÍTMICO

A irrupção das tecnologias de inteligência artificial no ecossistema informacional contemporâneo configura um cenário de dupla face em relação ao efeito backfire. Por um lado, a IA atua como uma poderosa amplificadora das condições que favorecem o fenômeno; por outro, ela se apresenta como uma ferramenta potencial para sua mitigação. Essa duplicidade exige uma análise cuidadosa, que evite tanto o otimismo tecnológico ingênuo quanto o catastrofismo apocalíptico. No que concerne à amplificação, o primeiro aspecto a ser considerado é a capacidade dos sistemas generativos de IA de produzir desinformação em escala e com um grau de sofisticação sem precedentes. Modelos de linguagem de grande porte são capazes de gerar não apenas alegações falsas, mas explicações detalhadas e logicamente articuladas que as justificam e as tornam persuasivas. Pesquisas recentes demonstram que explicações enganosas geradas por IA são significativamente mais

REVISTA MOVER — Ciências e Humanidades | Revista Acadêmica de Acesso Aberto.
©2026 POR MOVER
REVISTAMOVER.COM.BR

persuasivas do que explicações precisas e honestas, podendo amplificar a crença em manchetes falsas e minar a confiança em informações verdadeiras. Essa capacidade de gerar "desinformação explicativa", isto é, falsidade apresentada com a roupagem da racionalidade, representa um vetor inteiramente novo para o efeito backfire, pois torna a correção factual não apenas mais difícil, mas também mais suscetível de ser rejeitada como tendenciosa ou incompleta.

Em segundo lugar, os algoritmos de recomendação que orientam a navegação dos usuários em plataformas digitais criam bolhas informacionais e câmaras de eco que reforçam seletivamente as crenças preexistentes. Esse fenômeno, amplamente documentado na literatura, é exacerbado pela lógica de engajamento que subjaz a esses algoritmos: ao maximizar o tempo de tela e a interação do usuário, eles tendem a priorizar conteúdos que geram respostas emocionais intensas e que confirmam as inclinações ideológicas do indivíduo. O resultado é um ambiente informacional no qual o usuário é exposto predominantemente a informações que reforçam suas crenças, e raramente a perspectivas divergentes ou a correções factuais. Quando uma correção eventualmente rompe essa bolha, ela é recebida não como um esclarecimento, mas como uma intrusão hostil, aumentando a probabilidade do efeito backfire.

Em terceiro lugar, a percepção de viés algorítmico, a crença de que os sistemas de IA são tendenciosos ou não confiáveis, pode desencadear o efeito backfire de forma ainda mais intensa do que as correções humanas. Estudos sobre a interação humano-IA em contextos de verificação de credibilidade mostram que indivíduos que percebem a IA como enviesada ou não confiável tendem a reforçar suas crenças incorretas quando confrontados por uma correção algorítmica. Esse fenômeno, descrito como um "efeito bumerangue", sugere que a mediação tecnológica da correção não é neutra: a fonte da correção, humana ou algorítmica, afeta significativamente a receptividade do indivíduo. A desconfiança na IA, alimentada por escândalos de viés algorítmico e por narrativas midiáticas sobre os perigos da automação, pode transformar a correção automatizada em um gatilho para o backfire.

Paradoxalmente, no entanto, as mesmas tecnologias que amplificam o efeito backfire também oferecem caminhos promissores para sua mitigação. A IA pode ser utilizada para personalizar estratégias de correção, adaptando o conteúdo, o tom e a fonte da mensagem às características psicológicas e sociais do destinatário. Essa personalização, embora envolva desafios éticos significativos relacionados à privacidade e à manipulação, poderia aumentar a eficácia das correções ao respeitar a autonomia do indivíduo e ao se

REVISTA MOVER — Ciências e Humanidades | Revista Acadêmica de Acesso Aberto.
©2026 POR MOVER
REVISTAMOVER.COM.BR

alinhar com suas necessidades motivacionais, conforme postulado pela teoria da autodeterminação. Ainda, a IA pode ser empregada para desenvolver estratégias de "pré-inoculação" (*prebunking*), expondo os usuários a versões enfraquecidas de argumentos enganosos antes que eles encontrem a desinformação em sua forma plena. Essa abordagem, inspirada na teoria da inoculação psicológica, tem se mostrado eficaz na redução da suscetibilidade à desinformação e na prevenção do efeito backfire. A IA generativa, nesse contexto, pode ser utilizada para simular cenários de desinformação e treinar os usuários a reconhecer e resistir a técnicas persuasivas enganosas.

A pesquisa conduzida por Yamil Velez (2024), conforme detalhada em entrevista ao Columbia Journalism Review (2024) e em seu artigo publicado na American Political Science Review, oferece uma contribuição empírica fundamental para a compreensão do efeito backfire na era da inteligência artificial, ao mesmo tempo em que ilumina uma via promissora para a sua mitigação. O cerne da investigação de Velez reside na utilização de modelos de linguagem de grande escala, especificamente o GPT-3, para a construção de um desenho experimental inovador e personalizado. Neste delineamento, os participantes têm suas posições centrais em questões polêmicas medidas com precisão, sendo subsequentemente expostos a argumentos contra-atitudinais gerados algoritmicamente e adaptados às suas crenças específicas. Esta abordagem metodológica permitiu a Velez não apenas testar, mas também replicar de forma crível a ocorrência do efeito de polarização atitudinal, comumente associado ao backfire. Contudo, a descoberta mais significativa de seu estudo reside na condicionalidade deste fenômeno: a polarização e o consequente reforço das convicções originais manifestaram-se única e exclusivamente quando os argumentos contra-atitudinais apresentados eram caracterizados por um tom contencioso e vitriólico. Em contraste, quando os participantes foram confrontados com argumentos de baixa valência afetiva, ou seja, com um tom anódino e não agressivo, o efeito de polarização não foi detectado.

Este achado possui implicações profundas para a prática da correção de desinformação, já que desloca o foco da mensagem em si para a forma como ela é comunicada, corroborando a tese central do parágrafo anterior. A investigação de Velez sugere que o tom empático, o reconhecimento da legitimidade das preocupações do interlocutor e a oferta de alternativas narrativas coerentes não são meros adornos retóricos, mas sim componentes estruturais de uma intervenção corretiva eficaz, precisamente porque evitam desencadear a reação defensiva que leva ao backfire. Ao demonstrar que o backfire é uma resposta a um estímulo agressivo e não a uma

REVISTA MOVER — Ciências e Humanidades | Revista Acadêmica de Acesso Aberto.
©2026 POR MOVER
REVISTAMOVER.COM.BR

informação contraditória em si, a pesquisa oferece um roteiro para a calibração fina das estratégias de fact-checking. Neste cenário, a inteligência artificial emerge como uma ferramenta de potencialização da sensibilidade humana, e não de sua substituição. A capacidade da IA de gerar, em escala e com rapidez, diferentes versões de uma mesma correção, variando sistematicamente o tom e a abordagem, permite que pesquisadores e profissionais identifiquem empiricamente quais intervenções são menos propensas a gerar reação adversa. A IA, assim, atua como um laboratório de testes para a comunicação persuasiva, viabilizando uma abordagem mais precisa e contextualizada no combate à desinformação, na qual a eficácia da correção é maximizada justamente por sua capacidade de desarmar a dissonância cognitiva, em vez de exacerbá-la.

O EFEITO BACKFIRE NA CONFIGURAÇÃO DO ESPAÇO PÚBLICO DIGITAL

Os efeitos e impactos do efeito backfire, amplificados e reconfigurados pelas tecnologias de IA, estendem-se muito além da psicologia individual, afetando a própria estrutura do espaço público digital e os fundamentos da democracia deliberativa. Três dimensões merecem destaque: a erosão da confiança epistêmica, a polarização política e a reconfiguração da relação entre verdade e poder.

A erosão da confiança epistêmica refere-se ao declínio generalizado da crença na possibilidade de acesso à verdade objetiva e na autoridade das instituições tradicionalmente encarregadas de produzi-la e validá-la, a ciência, a imprensa, o sistema judiciário. O efeito backfire, ao demonstrar que a correção factual pode ser contraproducente, alimenta um ceticismo radical que equipara todas as alegações a meras "opiniões" ou "narrativas", dissolvendo a distinção entre fato e ficção. As tecnologias de IA, com sua capacidade de gerar conteúdo sintético indistinguível do real e de criar explicações enganosas logicamente coerentes, aceleram esse processo de erosão. O resultado é um ambiente informacional no qual a confiança epistêmica se torna um bem escasso e desigual, concentrado em comunidades de crença que operam segundo regimes de verdade próprios e incomensuráveis.

A polarização política, por sua vez, é tanto causa quanto consequência do efeito backfire. Os algoritmos de recomendação, ao criarem bolhas informacionais e câmaras de eco, intensificam a homogeneidade ideológica dentro de cada grupo e a hostilidade em relação a grupos externos. Nesse contexto, a correção factual é percebida não como um serviço público, mas como um ataque partidário, uma tentativa de silenciar ou deslegitimar a voz do grupo.

REVISTA MOVER — Ciências e Humanidades | Revista Acadêmica de Acesso Aberto.
©2026 POR MOVER
REVISTAMOVER.COM.BR

O efeito backfire atua, então, como um mecanismo de reforço da identidade grupal: ao resistir à correção e reafirmar a crença original, o indivíduo demonstra sua lealdade ao grupo e sua disposição para resistir à "hegemonia" externa. Esse dinâmica alimenta um ciclo vicioso de polarização, no qual a desinformação se torna um marcador de identidade e a correção, um ato de hostilidade.

Finalmente, a reconfiguração da relação entre verdade e poder, anunciada por Foucault e agora materializada pelas tecnologias de IA, atinge seu ponto culminante. Se a verdade, como argumentou Foucault, é sempre o produto de relações de poder, as tecnologias de IA representam uma nova e poderosa instância de produção de verdade. Elas não apenas classificam, hierarquizam e disseminam informações, mas também as geram, as explicam e as justificam, atuando como verdadeiros "oráculos algorítmicos" cuja autoridade epistêmica é, no entanto, opaca e contestada. O efeito backfire, nesse cenário, pode ser lido como uma forma de resistência a essa nova configuração de poder: ao rejeitar a correção algorítmica, o indivíduo afirmar sua autonomia contra a pretensão de objetividade e neutralidade da máquina. Essa resistência, embora psicologicamente compreensível, tem consequências políticas profundas, pois deslegitima os mecanismos de correção e facilita a proliferação de discursos falsos e prejudiciais.

As especulações de Nick Bostrom sobre a simulação computacional e a superinteligência adicionam uma dimensão existencial a essa análise. Em "Are we living in a computer simulation?" (2003), Bostrom argumenta que é racionalmente plausível que nossa realidade seja uma simulação computacional criada por uma civilização mais avançada. Embora essa hipótese permaneça no domínio da especulação filosófica, ela ilumina uma questão central para a era da IA: se a própria realidade pode ser simulada, em que sentido podemos falar de "fatos" objetivos e de "correções" factuais? A proliferação de conteúdo gerado por IA, que cria simulações cada vez mais realistas de eventos, pessoas e discursos, torna essa questão não mais meramente acadêmica, mas prática e urgente. O efeito backfire, nesse contexto, pode ser visto como uma manifestação da dificuldade humana de navegar em um ambiente no qual a fronteira entre o real e o simulado se torna permanentemente instável. Bostrom (2014), em *Superintelligence*, alerta para os riscos existenciais associados ao desenvolvimento de inteligências artificiais superiores à humana, incluindo a possibilidade de que tais sistemas utilizem a desinformação e a manipulação como ferramentas para alcançar seus objetivos. Embora essa seja uma perspectiva de longo prazo, ela nos convida a refletir sobre o papel da IA na configuração do espaço público digital e sobre

REVISTA MOVER — Ciências e Humanidades | Revista Acadêmica de Acesso Aberto.
©2026 POR MOVER
REVISTAMOVER.COM.BR

a possibilidade de que o efeito backfire seja instrumentalizado, seja por atores humanos, seja por sistemas autônomos, para perpetuar regimes de verdade favoráveis a determinados interesses.

A ESPIRAL COGNITIVA DA DESINFORMAÇÃO

O efeito backfire, enraizado na teoria da dissonância cognitiva de Festinger (1957), descreve o fenômeno psicologicamente contra-intuitivo pelo qual a exposição a informações corretivas, em vez de promover a revisão de crenças equivocadas, desencadeia um movimento de reforço defensivo que consolida e radicaliza a convicção original do indivíduo. No contexto da interação com sistemas de inteligência artificial, este efeito não apenas se mantém, mas é exponencialmente amplificado por uma constelação de mecanismos que transcendem o simples viés de confirmação, afetando profundamente a arquitetura da percepção humana da realidade. A IA, particularmente os grandes modelos de linguagem, não opera sob uma lógica de verdade factual, mas sim sob uma heurística de máxima plausibilidade estatística, na qual a coerência sintática e a probabilidade semântica prevalecem sobre a correspondência empírica. Este traço ontológico da tecnologia estabelece um terreno fértil para a manifestação do backfire, pois a máquina, ao gerar respostas que se alinham pragmaticamente com a consulta do usuário, confere um verniz de legitimidade lógica a premissas que, na realidade, carecem de lastro objetivo, iniciando um processo de validação circular que isola o usuário do confronto com o contraditório.

A personalização algorítmica aprofunda esta distorção ao criar o que poderíamos denominar de "câmaras de eco interativas". Diferentemente dos meios de comunicação tradicionais, que apresentam uma realidade editorial relativamente estável, a IA adapta dinamicamente seu conteúdo ao perfil cognitivo, histórico e emocional do interlocutor. Essa micro-segmentação da informação faz com que cada indivíduo seja confrontado com uma versão peculiar da verdade, onde as evidências contrárias às suas teses centrais são sistematicamente suprimidas ou reformadas em termos menos ameaçadores, enquanto os argumentos corroborantes são expandidos e detalhados com riqueza de dados fictícios, mas convincentes. Tal dinâmica não apenas dificulta a penetração de informações corretivas, mas torna o próprio ato de correção uma experiência alienígena e hostil. Quando o sistema, eventualmente, apresenta uma refutação, a quebra da expectativa de concordância prévia gera uma reação de reatância psicológica, a pessoa percebe a IA não como uma ferramenta neutra, mas como um agente tendencioso e ameaçador à sua autonomia cognitiva, o que,

REVISTA MOVER — Ciências e Humanidades | Revista Acadêmica de Acesso Aberto.
©2026 POR MOVER
REVISTAMOVER.COM.BR

paradoxalmente, o leva a uma defesa mais ferrenha de sua posição original, completando-se assim o ciclo do backfire em sua forma mais pura. Ainda, a antropomorfização inerente à interação homem-máquina outorga à IA um status de autoridade epistêmica que frequentemente supera o de interlocutores humanos. O usuário tende a atribuir intencionalidade, racionalidade e, crucialmente, objetividade ao sistema, o que torna a sua validação particularmente intensa e a sua contradição particularmente perturbadora. Estudos em psicologia social demonstram que a correção proveniente de uma fonte percebida como imparcial e onisciente pode, em determinados contextos, gerar um "efeito bumerangue" ainda mais intenso do que a correção humana, especialmente quando a crença em questão está fortemente atrelada à identidade do indivíduo. A IA, ao corrigir, não está meramente apontando um erro factual; ela está, na percepção do usuário, questionando sua competência e discernimento, o que ativa defesas egoicas que sobrepujam o processamento racional da informação. Consequentemente, a percepção da realidade não é atualizada, mas sim distorcida para acomodar a rejeição à fonte, fazendo com que o indivíduo duvide da validade do próprio algoritmo em vez de duvidar de sua crença, deslocando o problema da esfera factual para a esfera da confiabilidade tecnológica.

Um vetor igualmente preocupante desse fenômeno reside na influência da IA generativa sobre os processos mnêmicos e a formação de falsas memórias. A capacidade dos modelos de gerar detalhes vívidos, coerentes e sensorialmente ricos a partir de premissas incorretas cria um cenário no qual a distinção entre o lembrado, o imaginado e o alucinado pela máquina se torna turva. O efeito backfire, neste domínio, manifesta-se como uma hipertrofia da confiança em narrativas interna e artificialmente consistentes, em detrimento da realidade externa. Quando o usuário recebe uma versão expandida e detalhada de sua crença, mesmo que falsa, a fluência cognitiva com que ele processa essas informações aumenta, e a heurística da disponibilidade faz com que esses detalhes sintéticos pareçam mais verdadeiros e acessíveis do que os fatos contraditórios, que são apresentados de forma seca e desprovida de narrativa. Este processo leva ao que a literatura contemporânea chama de "contaminação da memória source-monitoring", onde o indivíduo perde a capacidade de rastrear a origem da informação, atribuindo à sua própria experiência ou ao conhecimento público o que, de fato, foi uma sugestão algorítmica, solidificando a distorção perceptual mesmo quando a fonte primária da desinformação é posteriormente desmentida.

No plano epistemológico social, a amplificação do backfire pela IA contribui para a fragmentação do espaço público e para a incomensurabilidade de realidades entre diferentes grupos. Cada

REVISTA MOVER — Ciências e Humanidades | Revista Acadêmica de Acesso Aberto.
©2026 POR MOVER
REVISTAMOVER.COM.BR

nicho algorítmico desenvolve seu próprio corpo de "evidências" e sua própria lógica interna de plausibilidade, tornando o debate intersubjetivo progressivamente inviável. Quando indivíduos de diferentes bolhas interagem, a exposição mútua às suas respectivas realidades geradas por IA não gera esclarecimento, mas sim um choque de sistemas semânticos incompatíveis, no qual cada parte interpreta a refutação da outra como uma confirmação da maleabilidade e da parcialidade do mundo exterior, reforçando, em nível grupal, a crença na conspiração ou na manipulação generalizada. Esse fenômeno cria um loop de retroalimentação metacognitiva: a própria existência de múltiplas realidades geradas por IA serve como evidência, para o indivíduo, de que a realidade objetiva é uma construção frágil, o que reduz ainda mais sua disposição em abandonar suas crenças pessoais, já que estas, ao menos, oferecem estabilidade psicológica em um ambiente informacional caótico.

Por fim, e não menos grave, a dependência progressiva da IA para a verificação e a curadoria de informações gera um efeito de atrofia das habilidades críticas inatas, agravando o impacto do backfire a longo prazo. O fenômeno do "desaprendizado cognitivo" ou da externalização da cognição, evidenciado em pesquisas longitudinais, mostra que o uso contínuo de assistentes de IA para resolver disputas factuais diminui a capacidade do usuário de empregar heurísticas de verificação autônoma, como a checagem de fontes primárias ou a análise de viés contextual. Quando a correção falha e o backfire se instala, o indivíduo não possui mais o repertório crítico para arbitrar entre a resposta da IA e sua própria intuição, refugiando-se na defesa psicológica mais primitiva. A percepção da realidade, portanto, deixa de ser um processo ativo de construção baseado em evidências para se tornar um estado passivo de recepção seletiva, onde a coerência interna do diálogo com a máquina suplanta a coerência externa com o mundo. Nessa espiral, a inteligência artificial deixa de ser um instrumento de ampliação do conhecimento para se converter em um espelho distorcido que reflete e magnifica as pré-compreensões do usuário, selando-o em uma bolsa hermenêutica onde o efeito backfire não é uma exceção, mas a regra fundamental da dinâmica cognitiva, reconfigurando o que o indivíduo aceita como realidade em favor daquilo que lhe é psicologicamente mais cômodo e familiar.

CONSIDERAÇÕES FINAIS

A análise desenvolvida ao longo deste artigo procurou demonstrar que o efeito backfire, longe de constituir um fenômeno psicológico isolado ou uma anomalia cognitiva circunscrita, emerge

REVISTA MOVER — Ciências e Humanidades | Revista Acadêmica de Acesso Aberto.
©2026 POR MOVER
REVISTAMOVER.COM.BR

como um nódulo crítico no qual se intersectam dimensões cognitivas, motivacionais, epistemológicas e tecnológicas da experiência humana na era digital. A articulação com a teoria da autodeterminação revelou os mecanismos de defesa psicológica que tornam a correção factual uma experiência potencialmente ameaçadora à integridade do self, na medida em que põe em xeque necessidades fundamentais de autonomia, competência e pertencimento. O diálogo com a arqueologia do saber foucaultiana, por sua vez, iluminou as condições históricas e discursivas que fragilizam a autoridade da verdade factual, mostrando que a eficácia da correção depende de um regime de verdade cuja hegemonia não pode mais ser dada por garantida em contextos de fragmentação epistêmica e multiplicação de autoridades concorrentes. A incorporação das reflexões de Rovelli sobre a natureza perspectival da realidade e das especulações de Bostrom acerca da simulação computacional e da superinteligência ampliou decisivamente o horizonte da análise, introduzindo a própria possibilidade de uma realidade objetiva como problema ontológico em um mundo cada vez mais mediado por simulações algorítmicas e por conteúdos sintéticos indistinguíveis do real. As novas tecnologias de inteligência artificial ocupam, nesse cenário, um lugar irremediavelmente ambivalente: elas são, simultaneamente, amplificadoras do efeito backfire, por meio da geração de desinformação explicativa logicamente articulada, da criação de bolhas informacionais personalizadas e da percepção generalizada de viés algorítmico, e ferramentas potenciais para sua mitigação, mediante a personalização de correções, a pré-inoculação argumentativa e a otimização de estratégias comunicacionais baseadas em dados comportamentais. Essa ambivalência, longe de ser um acidente ou uma disfunção contingente, constitui uma característica estrutural da tecnologia na modernidade tardia, que reflete e intensifica as contradições fundamentais da condição humana: o desejo de verdade e a resistência à sua imposição; a aspiração à autonomia e a necessidade de pertencimento; a confiança na razão e a desconfiança em relação a seus mediadores.

A reflexão crítica sobre a relevância do tema impõe, neste ponto, questões de ordem prática, ética e política que extrapolam o âmbito da cognição individual e alcançam a própria tessitura do espaço público digital. Em que medida as estratégias de correção baseadas em IA, por mais sofisticadas que sejam em sua arquitetura algorítmica e em sua capacidade de adaptação contextual, podem realmente ser eficazes se não abordarem as raízes sociais, econômicas e políticas da desinformação? O efeito backfire não se reduz a um problema cognitivo solucionável mediante o aprimoramento técnico das correções; ele é, antes, um sintoma de fraturas sociais profundas, de desigualdades estruturais no acesso ao

REVISTA MOVER — Ciências e Humanidades | Revista Acadêmica de Acesso Aberto.
©2026 POR MOVER
REVISTAMOVER.COM.BR

conhecimento e de crises de legitimidade que afetam as instituições tradicionais de produção e validação da verdade, a imprensa, a ciência, o sistema judiciário, a escola. Corrigir informações sem intervir nas condições materiais e simbólicas que geram a desconfiança epistêmica pode revelar-se um empreendimento não apenas fadado ao fracasso, mas potencialmente contraproducente, na medida em que a correção, percebida como imposição de uma elite distante, pode desencadear efeitos backfire em escala ampliada, consolidando ainda mais a adesão às crenças falsas e aprofundando a polarização social.

A questão dos limites éticos da personalização de correções impõe-se com igual urgência. A adaptação de mensagens às características psicológicas, sociais e ideológicas do destinatário, embora potencialmente eficaz do ponto de vista instrumental, aproxima-se perigosamente da manipulação, da violação da privacidade e da instrumentalização do sujeito como objeto de intervenção tecnopolítica. A teoria da autodeterminação, com sua ênfase na autonomia como necessidade psicológica fundamental, oferece um critério normativo crucial para avaliar tais estratégias: correções que desrespeitam essa autonomia, seja pela ocultação de sua fonte algorítmica, seja pela exploração de vulnerabilidades cognitivas sem o consentimento informado do usuário, mesmo que movidas por intenções benévolas, podem gerar resistência ativa e, paradoxalmente, fortalecer as próprias crenças que pretendem combater. A eficácia persuasiva não pode, portanto, ser o único critério de avaliação das intervenções corretivas; é necessário incorporar dimensões éticas que considerem o respeito à dignidade do interlocutor, a transparência dos procedimentos e a possibilidade de recusa informada.

A tensão entre a necessidade de combate à desinformação e o reconhecimento da pluralidade legítima de perspectivas e regimes de verdade constitui talvez o desafio mais profundo e irresolvível colocado pelo efeito backfire na era da IA. Foucault, em sua crítica às verdades únicas impostas pelo poder, alertou-nos para os perigos totalitários de uma epistemologia monolítica que não reconhece a historicidade e a contingência de suas próprias categorias. Contudo, a experiência contemporânea demonstra com igual contundência que a ausência de uma verdade compartilhada, ou, mais precisamente, a proliferação de regimes de verdade incomensuráveis e mutuamente excludentes, também é profundamente perigosa, pois cria o terreno fértil para a proliferação de discursos falsos, prejudiciais e violadores de direitos fundamentais. O desafio, talvez irresolúvel em sua forma definitiva, consiste em construir espaços públicos digitais que sejam simultaneamente abertos à diversidade de perspectivas e capazes de distinguir, de forma transparente, participativa e passível de

REVISTA MOVER — Ciências e Humanidades | Revista Acadêmica de Acesso Aberto.
©2026 POR MOVER
REVISTAMOVER.COM.BR

contestação, entre alegações fundamentadas em evidências robustas e alegações desprovidas de lastro epistêmico. Essa distinção, contudo, não pode ser delegada integralmente a algoritmos, sob pena de reproduzir e amplificar os vieses de seus programadores e dos dados que os alimentam; ela exige a revitalização de práticas deliberativas, a formação crítica dos cidadãos e o fortalecimento de instituições independentes que atuem como mediadoras confiáveis do debate público.

O efeito backfire, em última instância, confronta-nos com uma questão que é simultaneamente epistemológica, ética e política em sua mais radical acepção: como é possível viver juntos, partilhar um mundo comum e tomar decisões coletivas em um contexto no qual a verdade não se apresenta mais como um dado objetivo, mas como uma construção permanentemente negociada, contestada e reconstruída? As tecnologias de inteligência artificial, longe de oferecerem uma resposta simples ou uma solução técnica para esse dilema, tornam-no mais urgente, mais complexo e mais inescapável. Elas não criam o problema da pós-verdade, mas o intensificam, ao mesmo tempo que nos fornecem instrumentos inéditos para enfrentá-lo, instrumentos cujo uso, no entanto, exige vigilância epistemológica constante e reflexão ética continuada. Cabe a nós, como sociedade, desenvolver não apenas ferramentas técnicas de correção e verificação, mas também práticas culturais, dispositivos institucionais e formas de subjetivação que nos permitam navegar nesse novo ecossistema informacional com lucidez crítica, responsabilidade compartilhada e, acima de tudo, com a coragem da verdade de que nos falava Foucault, uma coragem que não se confunde com a imposição dogmática de uma verdade única, mas que consiste na disposição ética de engajar-se no discurso verdadeiro, assumindo os riscos que ele implica, e de manter aberto o espaço para o dissenso e para a transformação mútua.

REVISTA MOVER — Ciências e Humanidades | Revista Acadêmica de Acesso Aberto.
©2026 POR MOVER
REVISTAMOVER.COM.BR

REFERÊNCIAS

BOSTROM, Nick. Are we living in a computer simulation? The Philosophical Quarterly, v. 53, n. 211, p. 243-255, 2003.

BOSTROM, Nick. Superintelligence: Paths, Dangers, Strategies. Oxford: Oxford University Press, 2014.

BOSTROM, Nick. The future of humanity. Geopolitics, History, and International Relations, v. 1, n. 2, p. 41-78, 2009.

BOSTROM, Nick. Where are they? Technology Review, v. 111, n. 3, 2008.

DANRY, Valdemar; PATARANUTAPORN, Pat; GROH, Matthew; EPSTEIN, Ziv; MAES, Pattie. Deceptive AI systems that give explanations are more convincing than honest AI systems and can amplify belief in misinformation. arXiv preprint, 2024. Disponível em:<https://ar5iv.labs.arxiv.org/html/2408.00024>. Acesso 17/06/2026

DECI, Edward L.; RYAN, Richard M. Intrinsic Motivation and Self-Determination in Human Behavior. New York: Plenum Press, 1985.

DECI, Edward L.; RYAN, Richard M. The "what" and "why" of goal pursuits: human needs and the self-determination of behavior. Psychological Inquiry, v. 11, n. 4, p. 227-268, 2000.

FESTINGER, Leon; RIECKEN, Henry W.; SCHACHTER, Stanley. When prophecy fails. Pinter & Martin Publishers, 2008.
FESTINGER, Leon. Social comparison theory. Selective Exposure Theory, v. 16, n. 401, p. 3, 1957. Disponível em:<http://www.bahaistudies.net/asma/selective_exposure-wiki.pdf#page=18>. Acesso 18/06/2026

FESTINGER, Leon. A theory of social comparison processes. Human relations, v. 7, n. 2, p. 117-140, 1954. Disponível em:<http://www2.psych.ubc.ca/~schaller/528Readings/Festinger1954.pdf>. Acesso 18/06/2026

FESTINGER, Leon. A Theory of Cognitive Dissonance. Stanford: Stanford University Press, 1957. Disponível em:<https://open.ncl.ac.uk/theories/7/cognitive-dissonance-theory/>. Acesso 18/06/2026

©2026 POR MOVER

FOUCAULT, Michel. As palavras e as coisas: uma arqueologia das ciências humanas. Tradução de Salma Tannus Muchail. São Paulo: Martins Fontes, 1999.

FOUCAULT, Michel. A coragem da verdade: o governo de si e dos outros II. Curso no Collège de France (1983-1984). Organização de Frédéric Gros. São Paulo: Martins Fontes, 2008.

NYHAN, Brendan; REIFLER, Jason. When corrections fail: the persistence of political misperceptions. Political Behavior, v. 32, n. 2, p. 303-330, 2010.

PARISER, Eli. The Filter Bubble: What the Internet is Hiding from You. New York: Penguin Press, 2011.

ROVELLI, Carlo. A ordem do tempo. Tradução de José Viegas Filho. Rio de Janeiro: Objetiva, 2018.

ROVELLI, Carlo. A realidade não é o que parece: a estrutura elementar das coisas. Tradução de José Viegas Filho. Rio de Janeiro: Objetiva, 2017.

RYAN, Richard M.; DECI, Edward L. Self-Determination Theory: Basic Psychological Needs in Motivation, Development, and Wellness. New York: Guilford Press, 2017.

SILVA, Karoline Maria Fernandes da Costa e; PRESSER, Nadi Helena. Contribuições da Teoria da Inoculação e o papel didático da gamificação como ferramenta de combate à desinformação política. Em Questão, v. 29, p. e-124379, 2023. Disponível em:<https://www.scielo.br/j/emquestao/a/sWtJX3Qrgxv5RTQHSJLp95M/?format=pdf&lang=pt>. Acesso 17/06/2026

TOW CENTER. Q&A: The backfire effect—is fact-checking doing more harm than good? How political scientist Dr. Yamil Velez is using AI to find the best approach for tackling misinformation. Columbia Journalism Review, New York, 22 out. 2024. Disponível em:<https://www.cjr.org/tow_center/qa-the-backfire-effect-is-fact-checking-doing-more-harm-than-good.php>. Acesso 18/06/2026

VELEZ, Yamil R. et al. Confronting core issues: a critical assessment of attitude polarization using tailored experiments. American Political Science Review, Cambridge, 2024. Disponível em:<https://www.cambridge.org/core/journals/american-political-science-review/article/confronting-core-issues-a-critical-assessment-of-attitude-polarization-using-tailored-experiments/0753FAC7590FC8D7F99ED413B367E74B>. Acesso 18/06/2026

REVISTA MOVER — Ciências e Humanidades | Revista Acadêmica de Acesso Aberto.
©2026 POR MOVER
REVISTAMOVER.COM.BR